

## EPIWORK WP3 Second Year Report

### 1. Publishable summary

#### *Project objectives for the period*

*Please provide an overview of the project objectives for the reporting period in question, as included in Annex I of the Grant Agreement.*

*Please include a summary of the recommendations from the previous reviews (if any) and indicate how these have been taken into account.*

We are developing an information platform to mediate access to distributed collections of public health data, offering an easy and safe way to share data for those data providers who want to collaborate with epidemiological modellers. Researchers will use this platform in multiple ways: i) as catalogue of data sources containing the metadata describing existing databases; ii) as a forum to publish information about their own data, seeking modellers to collaborate with, and/or to seek sources of data that could be of interest to their epidemiological modelling efforts; and finally, iii) as the host of mediating software that can automatically process queries for epidemiological data available from the information sources connected to the platform.

Work on the reporting period was focused on the prototyping of the infrastructure for the platform, characterizing epidemic data and developing a common meta-model for querying data of interest for epidemiological modelling. The platform mediates access to distributed collections of public health data, offering an easy and safe way to share data for those data providers who want to collaborate with epidemiological modellers through a publicly available Web interface.

Research work on the third year of the project will be focused on setting the access control framework and interlinking epidemic data using semantic web standards. In parallel, we will continue the development of a second version of the Epidemic Marketplace with much improved usability for curating epidemics data, which will also incorporate the new research results. The development of methods for assisting collaborators in easily find each other, and define mutual voluntary agreements for sharing their data will addressed mostly in the fourth year.

### 2.1 Background Information on WP3

Work package number	3	Start date or starting event:			Month 1			
Work package title	Information platform							
Activity type	RTD							
Participant number	12	1	2	3	4	5	9	10
Participant short name	FFCUL	ISI	FGC-IGC	TAU	MPI-DS	AIBV	BIU	FBK-IRST
Person-months per participant	82	60	4	8	42	2	11	6

This Work Package is lead by FFCUL. The whole WP3 activity is structured into four tasks:

### **Task 3.1 – Data Collection.**

**Participants:** FFCUL, ISI, FBK-IRST, BIU, MPI-DS, FGC-IGC, AIBV.

**Description:** Realistic simulations of epidemic processes crucially depend on the availability of datasets describing human behaviour and pathogen-host interactions. Datasets include population movement data, social and behavioural data, health related data, geographic data, detailed geo-temporal epidemic incidence and immunization data, pathogen evolution and multi-strains circulation data. Data can come from a variety of different sources, including hospital records, country statistics, Web content, and others. It can range from a global scale, such as the worldwide air transportation infrastructure, down to the detailed description of individual **activities** at a minute-by-minute scale. This task will create a catalogue of databases of epidemiological data across Europe, with extensive meta-data describing the main characteristics of the available information sources. This catalogue will be integrated with a collaborative platform that will be set up for online discussion and exchange of meta-data among the participants.

### **Task 3.2 – Meta-Model Design.**

**Participants:** FFCUL, IGC, ISI.

**Description:** While some of the previously mentioned datasets are freely available on the Web (e.g. WHO Global Health Atlas, Eurostat), they are often scattered in different repositories, cover partial regions of the world and come in different formats, according to different standards and classifications. The project envisions a unified and integrated approach for the management of these resources, with the design and implementation of an Epidemic Marketplace Platform, publicly available on the web. The platform supports the sharing and management of epidemic datasets and resources as well as their rating, annotation, and selection. It is an on-line social networking site that will serve researchers, practitioners, and educators all over the world to foster a virtual community for epidemic research. It will support the exchange of resources as well as user interactions. Based on a Web2.0 approach, users will become active participants, sharing information and data, and collaborating online, rather than being satisfied with a passive information consumer/viewer role. We envision proposing a simple reference format, which will facilitate the navigation and use of the datasets. Each dataset will come with a metadata file, signalling general

metadata for resource management, containing data such as: the title, the date of submission, version, the source of the data and coverage. Moreover, the metadata will include information for a more thorough description of the data included in the dataset, providing a framework for a more specific description, for example, of epidemiologic and geographic data,. The Marketplace will support flexible and intuitive tools for navigation and selection of resources. Standard classifications as well as tagging systems proposed by users will be supported.

### **Task 3.3 – Epidemic Marketplace Platform.**

**Participants:** FFCUL, ISI.

**Description:** This task will implement a platform based on the integration of grid technology and publicly available services and software on the web to support the sharing and management of epidemic datasets and resources as well as their rating, annotation, and selection. The Epidemic Marketplace Platform will be an on-line social networking site that will serve researchers, practitioners, and educators all over the world to foster a virtual community for epidemic research. It will support the exchange of resources as well as user interactions. Based on some of the Web2.0 characteristics, users will become active participants, generating information and providing data for sharing, and collaborating online, rather than being satisfied with a passive information consumer/viewer role. More specifically, researchers can use and contribute to the Marketplace in several different ways. They can: (1) use it as a catalogue of data sources containing the metadata describing existing databases; (2) view, download, tag, and comment on the available resources; (3) provide compliant datasets and relevant information; (4) use it as a forum where to publish information about their own data, seek modellers to collaborate with, share and distribute their new findings.

### **Task 3.4 – Evaluation and monitoring of the use of the catalogue and collaboration services.**

**Participant:** FFCUL.

**Description:** This task involves the monitoring of epidemiological data exchanges performed through the mediating services platform. The evaluation will assess not only the coverage of the catalogued resources, but the users' satisfaction with the user interface and integrated collaborative tools made available through the epidemiological marketplace platform. More importantly, the analysis of the collected datasets and their annotations and usage will provide a rich environment for deriving an epidemiology ontology, which will help further on the integration and communication among the community of epidemiologists.

## 2. WP Work progress and achievements during the period

- *A summary of progress towards objectives and details for each task;*
- *Highlight clearly significant results;*
- *Team publications within the cope of the project (please provide a pdf of the publication if possible);*
- *List of outreach activities (conferences, Invited talks, presentations, workshops, tutorials,);*
- *List of press releases or media coverage, any particular dissemination activity;*
- *unanticipated finding, opportunity etc.*
- *If applicable, explain the reasons for deviations from Annex I and their impact on other tasks as well as on available resources and planning;*
- *If applicable, explain the reasons for failing to achieve critical objectives and/or not being on schedule and explain the impact on other tasks as well as on available resources and planning (the explanations should be coherent with the declaration by the project coordinator) ;*
- *a statement on the use of resources, in particular highlighting and explaining deviations between actual and planned person-months per work package and per beneficiary in Annex I (Description of Work)*
- *If applicable, propose corrective actions.*

## **2.2 Progress in the Reporting Period**

1. Involved partners attended the Epiwork Review in Brussels in March, 2010 – Mário Silva presented the two completed WP3 deliverables to the EC and gave an overview of the progress and challenges in WP3. The demo of the fully functional Epidemic Marketplace prototype was presented to the reviewers, who had the opportunity to obtain user accounts for browsing the EM repository and forum.
2. **[outreach]** WP3 members attended the COSI-ICT cluster workshop in Brussels, March, 2010.
3. **[opportunity]** Francisco Couto participated in the VPH-OBO ontology workshop at Cambridge UK where relevant case studies and guidelines in ontology development in the biomedical field were presented. During the workshop, Francisco discussed with Barry Smith (National Center for Ontological Research, USA) the state of the art of epidemiological ontologies and a possible collaboration within the IDO (infectious disease ontology).
4. **[outreach]** Mário Silva and Daniela Polotti attended WP5 Meeting in Amsterdam in May; Mário Silva and Daniela Paolotti lead the discussions on the Integration and privacy issues regarding collected data on WP3/WP5.
5. Major reorganization of the base software infrastructure of the Epidemic Marketplace, which is now entirely deployed in various virtual machines hosted mounted in the two dedicated servers and two network storage units.
6. Evaluation of Integration of Drupal with Fedora Commons, using the Islandora module extension to Drupal and using this software as basis for a new version of the Repository/Forum of the Epidemic Marketplace.
7. **[outreach]** Mário J. Silva and Daniela Paolotti attended the ASSYST Perada Workshop - Towards a Science of Socially Intelligent ICT in London, August 2010, where he made the case for handling privacy from the ground-up in information systems collecting and managing social networking data (such as Epiwork).

8. **[outreach]** WP3 participants attended ECCS'10, the European Conference on Complex Systems in Lisbon, September 2010 and the COSI-ICT workshop.
9. WP3/WP4 meeting in Lisbon on September 16 at ECCS'10. The contents of the EM were discussed with Daniela Paolotti, who submitted the data files used by the GLEaMviz simulator, with focus on appropriate metadata for the categorization of the resources. Some concepts behind the development of EM v2.0 were discussed, with focus the user-oriented interface. There was some brainstorming on the integration between WP3 and WP4 was held. An initial approach to integration was discussed which consisted of a Client Application that checks the EM for updates to the subscribed collections and downloads them to the WP4 server. GLEaMviz would then run simulations using the data and the Client will upload the Flash Visualization Files back to the Epidemic Marketplace. This was the idea behind the SimpleEMClient that we developed later.
10. **[significant result]** Public Release of the Epidemic Marketplace (Deliverable 3.3) at <http://epimarketplace.net>. Anyone from the community can now define collections and retrieve/upload datasets.
11. WP3 participants attended the Second Epiwork Meeting in Torino, December 2010. Mário Silva presented the status of WP3 and João Zamite introduced Privacy + Access Control and the plans of handling the issue at the Epidemic Marketplace.
12. In a WP3/WP4 meeting in Torino December the WP3 and WP4 members discussed plans for further integration of the data and computational platforms. In a new scenario to be demonstrated using the EM platform and the GLEaMviz simulator, simulation files will be deposited in the EM, then sent to the simulation platform and finally the results will be transferred from the computation platform to the Epidemic Marketplace. The simulation parameters and the results would be shared by the users and computational tools having appropriate access permissions.
13. **[opportunity]** Initiated design of new access control method to be implemented in the EM, whose users will be able to specify user access permissions based on their Social Networks membership. This has enormous potential for simplifying the process of managing access permissions. These ideas were initially presented at the 2<sup>nd</sup> Epiwork Meeting in Torino, in December 2010.

14. **[publication]** In January, Zamite et al. submitted an extended version of the paper “MEDCollector: Multisource Epidemic Data Collector” to a special issue of the Springer journal LNCS Transactions on Large-Scale Data and Knowledge-Centered Systems entitled Database Systems for Bio-medical Applications.
15. SimpleEMClient available for download from the Epidemic Marketplace at the end of January, 2010.
16. EM 2.0 platform development underway for release in 2011. Major features: new user interface and component integration based on the Drupal Content-Management System, improved access control management, redesigned web-services.

### **Publications and Presentations:**

1. **[publication]** Mário J. Silva, Francisco M. Couto, Dulce Domingos, Juliana Duque, Hugo Ferreira, Luís F. Lopes, Daniela Paolotti, Fabrício Silva, Patrícia Sousa, João Zamite, D 3.3 Public Release of the Epidemic Marketplace Platform Technical Report. Technical Report . LASIGE, University of Lisbon, Faculty of Sciences, September 2010. Also available from the [epiwork.edu](http://epiwork.edu) website.
2. **[publication]** Luis Filipe Lopes, A Metadata Model for the Annotation of Epidemiological Data Master Thesis, University of Lisbon, Faculty of Sciences, September 2010.
3. **[publication]** João Zamite, Multisource Epidemic Data Collector Master Thesis, University of Lisbon, Faculty of Sciences, September 2010.
4. **[publication]** Hugo Ferreira, O Mediador do Epidemic Marketplace Master Thesis, University of Lisbon, Faculty of Sciences, January 2011.
5. **[publication]** Fabrício A.B. Silva, Mário J. Silva, Francisco Couto 2010: Epidemic Marketplace: an e-Science Platform for Epidemic Modelling and Analysis. ERCIM News 82(), 43-44. Special Theme: Computational Biology.
6. **[publication]** Fabrício A.B. Silva, Mário J. Silva, Francisco Couto 2010: Epidemic Marketplace: an e-Science Platform for Epidemic Modelling and Analysis. ERCIM News 82 - Special Theme: Computational Biology.

7. **[publication + presentation]** Luis Filipe Lopes, Fabrício A.B. Silva, Francisco Couto, João Zamite, Hugo Ferreira, Carla Sousa, Mário J. Silva, Epidemic Marketplace: An Information Management System for Epidemiological Data. Presented at ITBAM'10 - 1st International Conference on Information Technology in Bio- and Medical Informatics - DEXA 2010 - August, 2010.
8. **[publication + presentation]** João Zamite, Fabrício A.B. Silva, Francisco Couto, Mário J. Silva, MEDCollector: Multisource Epidemic Data Collector. Presented at ITBAM'10 - 1st International Conference on Information Technology in Bio- and Medical Informatics - DEXA 2010 - August, 2010.
9. **[publication + presentation]** Mário J. Silva, Fabrício A.B. Silva, Luís Filipe Lopes, Francisco Couto, Building a Digital Library for Epidemic Modelling. Proceedings of ICDL 2010 - The International Conference on Digital Libraries 1, p. 447--459, New Delhi, India, 23--27 February, 2010. TERI Press -- New Delhi, India. Presentation of invited paper.,
10. **[presentation]** Presentation and demo of EPIWORK to students at the University of Valencia, Spain in an invited 4hrs seminar, by Fabricio Silva, June 2010.
11. **[presentation]** Presentation of EPIWORK to students of the Master/Phd in Epidemiology at the Faculty of Medicine of the University of Lisbon, in an invited 3h seminar, by Mário J. Silva.
12. **[presentation]** Mário J. Silva, Privacy in Socially Intelligent ICT. Imperial College, London, UK. ASSYST Perada Workshop - Towards a Science of Socially Intelligent ICT. August, 2010.

### **Activities in the second year of the project in Task 3.1:**

The data collection activity started at the end of the first year. We have been automatically collecting the following data with the MedCollector since then, from the following sources:

- Messages from Twitter containing disease names and posted at each country or capital of the world.
- Google Flu Trends Datasets.



- RSS Feeds from the CDC. Including travel notices alerts and emerging diseases alerts.
- Email Newsletter messages from ProMEDmail.

These datasets are now being periodically assembled and uploaded into the Epidemic marketplace.

### **Activities in the second year of the project in Task 3.2:**

Work in this task included:

- Population of the repository with datasets and resources found relevant to epidemic modellers gathered from the Web, used for validation of the metadata modelling decisions.
- Development of a metadata editor for epidemic datasets for the EM platform, following the policies outlined in deliverable D3.1.

### **Activities in the second year of the project in Task 3.3:**

- Reconfiguration of the infrastructure, now based on virtual machines. Configuration of the production and development environments for the Lisbon EM node.
- Deployment of a new version of the EM base software, to be based in Fedora Commons version 3.0 and the Drupal Content Management System, initiated in January continued. This new version will substitute the current, based on Fedora Commons 2.2.2 and Muradora 1.3.3. This new version is planned to replace the production environment of the EM in 2011.
- The first operational prototype of the Epidemic Marketplace was presented in the March 2010 review in Brussels.
- Discussions with partners involved in WP3 on how to identify relevant datasets to the catalogue and strategies and incentives for populating the Epidemic Marketplace.
- WP3/WP4 collaboration on the integration of the computational platform with the data platform.
- Public release of the Epidemic Marketplace at <http://epimarketplace.net>. Anyone from the community can now define collections and retrieve/upload datasets.

- Initiated design of new access control system to be implemented in the EM..
- Development of the SimpleEMClient, available for download from the Epidemic Marketplace.
- EM 2.0 platform development underway for release in 2011. Major features: new user interface and component integration based on the Drupal Content-Management System, improved access control management, redesigned web-services.

#### **Activities in second of the project in Task 3.4:**

- Continued work on planning of the monitoring and log data collection and analysis tasks. Collection of usage data is now active since the release of the Epidemic Marketplace to the public. This is done through Google analytics and analysis of the web servers log data.

## **2.3 Outcast for WP3**

Focus for the 3rd year will be on providing the methods to further achieve INTEGRATION with the EPIWORK partner's systems and tools. In addition, we will be focusing on interlinking epidemic data in the Epidemic Marketplace using ontology engineering technologies (a deliverable on ontologies is scheduled at the end of the 3rd year). In parallel, we will continue developing the software of the Epidemic Marketplace, improving its interfaces and populating it with epidemic datasets.

The envisaged activities will mainly involve:

- Redesigning the user interface of the EM, and moving to having version 2.0 of the Epidemic Marketplace (Drupal/Islandora/Fedora Commons 3-based) in production.
- Continuing work on populating and documenting the EM Repository.
- Addressing access control for data sharing within the community of epidemiologists, and ethics, privacy and anonymization issues that must be tackled by the Epidemic Marketplace. Privacy and anonymization has been a subject of discussion with WP5 and within the whole project, because it must be tackled more comprehensively. For instance, Internet Monitoring Sources should anonymize datasets locally, instead of uploading them for later anonymization at the EM.

- Providing tools for selecting datasets or parts of the information thereof from the EM collection for retrieval access by epidemic modelling tools.
- Incorporating in the systems of project partners a Distributed Authentication and Access control mechanism.
- Review current work on epidemiological ontologies towards the development of ontological integrative approach that will help further on the integration and communication among the community of epidemiologists.
- Development of the first demonstrations of information search based on ontologic knowledge integrated in the Epidemic Marketplace.
- Revisiting proposed plans for adopting Grid Computing standards in the EM will in the third year. Cloud Computing became recently a core component of the offer of major IT companies and we are witnessing that platforms like Microsoft Azure have recently proposed “data marketplaces” whose concept has many similarities with the idea of an Epidemic Marketplace proposed by Epiwork. Given the embrace by the industry, the EM concept makes now more sense than ever and confirms that we have been pushing the development in the right direction, but we should also align it with the new industry offers in this domain.

### 3. Deliverables and milestones tables

#### Deliverables (excluding the periodic and final reports)

*Please list all the deliverables due in this reporting period, as indicated in Annex I of the Grant Agreement. Deliverables that are of a nature other than written "reports", such as "prototypes", "demonstrators" or "others", should also be accompanied by a short report, so that the European Commission has a record of their existence. If a deliverable has been cancelled or regrouped with another one, please indicate this in the column "Comments". If a new deliverable is proposed, please indicate this in the column "Comments".*

*This table is cumulative, that is, it should always show all deliverables from the beginning of the project.*

TABLE 1. DELIVERABLES <sup>1</sup>									
Del. no.	Deliverable name	WP no.	Lead beneficiary	Nature	Dissemination level	Delivery date from Annex I (proj month)	Delivered Yes/No	Actual / Forecast delivery date	Comments
3.1	Report - Meta-model initial specification, catalogue of relevant data, platform requirements.	3	FFCUL		Consortium	8	Yes	September, 30 <sup>th</sup> 2009	
3.2	Prototype of the Epidemic Marketplace Platform with an initial set of epidemiological databases integrated available to project participants	3	FFCUL		Consortium	12	Yes	January 31 <sup>st</sup> , 2010	

<sup>1</sup> For Security Projects the template for the deliverables list in Annex A1 has to be used.

3.3	Public Release of the Epidemic Marketplace Platform	3	FFCUL	Public	12	Yes	September, 30th 2010	

### Milestones

*Please complete this table if milestones are specified in Annex I of the Grant Agreement.  
Milestones will be assessed against the specific criteria and performance indicators as defined in Annex I.*

TABLE 2. MILESTONES								
Milestone no.	Milestone name	Work package no	Lead beneficiary	Delivery date from Annex I	Achieved Yes/No	Actual / Forecast achievement date	Comments	
5	Public release of Epidemic Marketplace Platform.	3	FFCUL	20	Yes	September, 30th 2010		